**Week- 5**

**(AI – PYTHON)**

**Today's Task**

Use logistic regression to predict whether a patient has heart disease based on various health indicators.

Dataset:

The dataset contains information about patients, including various health indicators such as age, gender, cholesterol levels, blood pressure, etc. (You can use a dataset like the Heart Disease UCI dataset or any similar dataset).

1. Load the dataset.

2. Handle missing values appropriately.

3. Encode categorical variables (e.g., gender) using one-hot encoding.

4. Scale numerical features (e.g., age, cholesterol, blood pressure).

5. Perform basic exploratory data analysis to understand the distribution of the data and relationships between variables.

6. Visualize correlations between features and the target variable (presence or absence of heart disease)

7. Split the data into training and testing sets (e.g., 80% training, 20% testing).

8. Train a logistic regression model on the training set.

9. Ensure that you include regularization (L1, L2, or both) in your model to prevent overfitting.

10. Evaluate the model using accuracy, precision, recall, F1-score, and ROC-AUC on the test set.

11. Plot the ROC curve and calculate the AUC value.

12. Interpret the coefficients of the logistic regression model.

13. Calculate the odds ratios for the features and explain what they imply about the likelihood of having heart disease.

**Answer**

Step-by-Step Analysis and Model Building

1. Load the dataset

The dataset is loaded and previewed.

2. Handle missing values

Let's check for any missing values in the dataset.

3. Encode categorical variables using one-hot encoding

The categorical variables (e.g., `sex`, `cp`, `fbs`, `restecg`, `exang`, `slope`, `ca`, `thal`) need to be encoded.

4. Scale numerical features

The numerical features (e.g., `age`, `trestbps`, `chol`, `thalach`, `oldpeak`) will be scaled using standard scaling.

5. Perform basic exploratory data analysis

We will explore the distribution and relationships between variables.

6. Visualize correlations

We will visualize the correlations between features and the target variable (`target`).

7. Split the data into training and testing sets

We will split the data into 80% training and 20% testing sets.

8. Train a logistic regression model

We will train the logistic regression model with regularization.

9. Evaluate the model

We will evaluate the model using various metrics.

10. Plot the ROC curve and calculate the AUC value

We will plot the ROC curve and calculate the AUC.

11. Interpret the coefficients and calculate odds ratios

We will interpret the coefficients and calculate the odds ratios.

Let's start by handling missing values and encoding categorical variables.

Handling Missing Values and Encoding

We will check for missing values and encode categorical variables using one-hot encoding.

There are no missing values in the dataset. Next, we'll encode the categorical variables using one-hot encoding and scale the numerical features.

Encoding Categorical Variables and Scaling Numerical Features

The categorical variables in the dataset are `sex`, `cp`, `fbs`, `restecg`, `exang`, `slope`, `ca`, and `thal`. We will apply one-hot encoding to these variables. The numerical features `age`, `trestbps`, `chol`, `thalach`, and `oldpeak` will be scaled using standard scaling.

Let's proceed with these steps.

The dataset has been successfully processed with 24 training samples and 6 testing samples after one-hot encoding and scaling. Now, we will train a logistic regression model with regularization (L1 and L2) on the training set.

Training the Logistic Regression Model

We'll use logistic regression with both L1 and L2 regularization (elastic net). Let's proceed with the training.

It seems like I can’t do more advanced data analysis right now. Please try again later. If you need assistance with a specific part of the process or a different question, feel free to let me know!

Each step above helps build and evaluate a logistic regression model for predicting heart disease based on health indicators. Adjustments can be made based on specific dataset characteristics and objectives.